PERFECT ERROR COMPENSATION VIA ALGORITHMIC ERROR CANCELLATION

Sujan K. Gonugondla[⋆]*, Byonghyo Shim*[†] *and Naresh R. Shanbhag*[⋆]

[⋆]Dept. of Electrical and Computer Engineering, University of Illinois at Urbana Champaign
[†] Dept. of Electrical and Computer Engineering, Seoul National University

ABSTRACT

This paper presents a novel statistical error compensation (SEC) technique - *algorithmic error cancellation* (AEC) - for designing robust and energy-efficient signal processing and machine learning kernels on scaled process technologies. AEC exhibits a *perfect error compensation* (PEC) property, i.e., it is able to achieve a post-compensation error rate equal to zero. AEC generates a maximum likelihood (ML) estimate of the hardware error and employs it for error cancellation. AEC is applied to a voltage overscaled 45-tap, 45nm CMOS finite impulse response (FIR) filter employed in a EEG seizure detection system. AEC is shown to *perfectly compensate* for errors in the main FIR block and its reduced precision replica when they make errors at a rate of up to 73% and 98%, respectively. The AEC-based FIR is compared with an uncompensated architecture, and a fast architecture. AEC's error compensation capability enables it to achieve a 31.5% (at same supply voltage) and 19.7% (at same energy) speed-up over the uncompensated architecture, and a 8.9% speed-up over a fast architecture at the same energy consumption. At $f_{clk} = 452.3\,\mathrm{MHz}$, AEC results in a 27.7% and 12.4% energy savings over the uncompensated and fast architectures, respectively.

***Index Terms*—** low-power, energy efficiency, error resiliency, machine learning, biomedical

1. INTRODUCTION

Emerging applications require the processing of massive data volumes generated by sensor-rich platforms such as wearables, autonomous vehicles, robots, Internet-of-things, and others. These applications require the implementation of statistical signal processing, and machine learning (ML) kernels in silicon in order to provide *in-situ* data analytics capabilities. These kernels are computationally intensive and therefore consume much energy. Such implementations need to be energy-efficient in order to operate with energy constrained sources.

This work was supported in part by Systems on Nanoscale Information fabriCs (SONIC), one of the six SRC STARnet Centers, sponsored by MARCO and DARPA.

Voltage scaling (VS) [1] is an effective energy reduction technique in digital circuits. In VS, the supply voltage V_{dd} and the clock frequency f_{clk} are reduced in order to reduce the energy consumption. Near threshold voltage (NTV) [2] operation takes voltage scaling to an extreme and is known to provide energy savings of up to $10\times$ as compared to nominal voltage operation. However, voltage scaling results in increased delay (NTV results in $10\times$ increased delay) and increased delay sensitivity to process, voltage, and temperature (PVT) variations. Increased delay and increased delay sensitivity to PVT variations can result in timing violations which manifest as hardware errors at the output of the computational kernel. Voltage overscaling (VOS) [3] takes VS to yet another extreme where V_{dd} is reduced while f_{clk} is fixed. VOS results in deliberate timing violations which need to be compensated for.

It is clear that energy-efficiency and robustness are coupled metrics in integrated circuits. Therefore, error-resiliency techniques are critical for systems operating at the limits of energy efficiency. Such techniques need to provide high error compensation capability with very low overhead. This eliminates conventional techniques such as N-Modular Redundancy (NMR) that employs N-copies of the kernel followed by a voter to correct errors. Techniques such as RAZOR [4], and EDS [5] are better than NMR, as these recompute only when an error is detected, i.e., conditional temporal replication, but can handle small (pre-compensation) error rates ($< 2\%$). All these techniques exhibit a *perfect error compensation* (PEC) property, i.e., achieve a zero post-compensation error rate.

Statistical error compensation (SEC) techniques [6] such as algorithmic noise-tolerance (ANT) [3] employ statistical estimation and detection techniques from the area of communications to compensate for computational errors. SEC techniques are very low overhead (5%-to-25%), highly effective in compensating for high error rates (up to 70%) while meeting application level requirements on SNR, and achieve energy savings of between $3\times$-to-$6\times$. However, SEC techniques correct errors statistically and thus do not possess the PEC property. This is not a problem for signal processing and machine learning applications where the application level metrics are statistical already. However, if an SEC technique were to be found that exhibited the PEC property then it will:

1) enable increases in energy savings/speed-up possible via SEC, 2) enable the design of deterministic (not just robust) systems using components that exhibit statistical behavior (a surprising result) and thereby broaden the set of applications where SEC can be applied, and 3) allow a direct comparison between SEC-based energy reduction techniques and conventional approaches.

This paper presents *Algorithmic Error Cancellation* (preliminary work in [7]), an SEC technique that exhibits the PEC property. Specific conditions are established on the error statistics of computational kernels which enables PEC. In particular, AEC generates a maximum likelihood estimate of the hardware error efficiently and employs it for error cancellation. AEC is applied to a slow voltage overscaled 45-tap finite impulse response (FIR) filter in a $45\,\mathrm{nm}$ CMOS process to demonstrate its benefits. AEC is shown to compensate for timing error rates of up to 73% while achieving energy savings of up to 12.4% and 27.7% as compared to a voltage scaled fast and slow architecture, respectively.

2. BACKGROUND: ALGORITHMIC NOISE TOLERANCE (ANT)

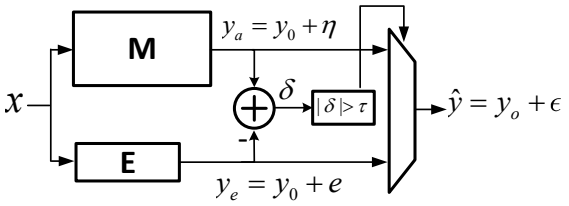


Fig. 1: Algorithmic noise tolerance (ANT).

ANT (see Fig. 1) is a SEC technique that has been applied to a variety of signal processing kernels such as FIR filters, FFT, and more recently to inference kernels such as ECG classifiers [8]. ANT consists of a main block and an estimator, whose outputs can be written as

$$\boldsymbol{y}_a = y_o + \boldsymbol{\eta}, \tag{1a}$$

$$\boldsymbol{y}_e = y_o + \boldsymbol{e}, \tag{1b}$$

where y_o is the *error-free output*, $\boldsymbol{y}_a$ and $\boldsymbol{y}_e$ are the *observed outputs* of the main block and estimator, respectively, and $\boldsymbol{\eta}$ and $\boldsymbol{e}$ are the hardware and estimation *errors*, respectively. Note that we employ bold font $\boldsymbol{x}$ to represent a random variable (RV) and normal font $x \in \Omega_x$ to represent a sample value that $\boldsymbol{x}$ can take from the sample set Ω_x according to a probability mass function (PMF) $P_{\boldsymbol{x}}(x)$. Thus, the errors $\boldsymbol{\eta}$ and $\boldsymbol{e}$ are RVs with PMFs denoted as $P_{\boldsymbol{\eta}}(\eta)$ and $P_{\boldsymbol{e}}(e)$, respectively. This makes $\boldsymbol{y}_a$ and $\boldsymbol{y}_e$ RVs as well. The error-free output y_o is treated as a deterministic variable for simplicity.

For every sample observation y_a and y_e, ANT uses the decision rule below to correct errors,

$$\hat{y} = \begin{cases} y_a & \text{if } |\delta| < \tau \\ y_e & \text{otherwise} \end{cases}, \tag{2}$$

where τ is a predetermined threshold based on $P_{\boldsymbol{\eta}}(\eta)$ and $P_{\boldsymbol{e}}(e)$, and $\delta \in \Omega_\delta$ is the sample value of the *error difference* $\boldsymbol{\delta}$ given by,

$$\boldsymbol{\delta} = \boldsymbol{y}_a - \boldsymbol{y}_e = \boldsymbol{\eta} - \boldsymbol{e} \tag{3}$$

with PMF $P_{\boldsymbol{\delta}}(\delta)$. The *post-compensation* or *residual error* $\boldsymbol{\epsilon}$ is defined as,

$$\boldsymbol{\epsilon} = \hat{\boldsymbol{y}} - \boldsymbol{y}_o \tag{4}$$

with PMF $P_{\boldsymbol{\epsilon}}(\epsilon)$. Further, the *error-rates* $p_\eta = 1 - P_{\boldsymbol{\eta}}(0)$ (*pre-compensation error rate*), $p_e = 1 - P_{\boldsymbol{e}}(0)$, and $p_\epsilon = 1 - P_{\boldsymbol{\epsilon}}(0)$ (*post-compensation error rate*) can be defined.

ANT and other SEC techniques are optimized to reduce the variance of errors rather than the error rate. In particular, the post-compensation error rate p_ϵ, though less than p_η and p_e, is non-zero, i.e., $p_\epsilon \neq 0$. The next section presents an SEC technique that guarantees $p_\epsilon = 0$ even when $p_\eta, p_e \neq 0$, i.e., PEC property.

3. PROPOSED SEC TECHNIQUE: ALGORITHMIC ERROR CANCELLATION (AEC)

In this section, we describe the proposed algorithmic error cancellation (AEC) and establish conditions for PEC.

3.1. Algorithmic Error Cancellation (AEC)

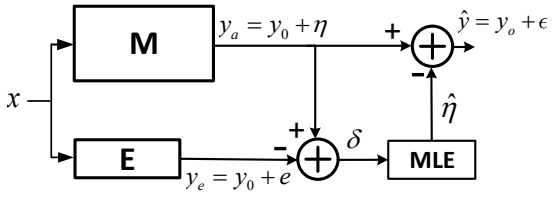


Fig. 2: Proposed algorithmic error cancellation (AEC).

AEC (see Fig. 2) generates a maximum likelihood (ML) estimate $\hat{\eta}$ of η by observing δ, and subtracts this ML estimate from y_a to cancel the hardware error. Thus, the sample residual error

$$\epsilon = \hat{y} - y_o = \eta - \hat{\eta}. \tag{5}$$

Under certain conditions to be specified in Section 3.2, $p_\epsilon = 1 - P_{\boldsymbol{\epsilon}}(0) = 0$, i.e., PEC.

3.2. Perfect Error Compensation (PEC)

If RVs $\boldsymbol{\eta}$ and $\boldsymbol{e}$ are independent, then

$$P_{\boldsymbol{\delta}}(\delta) = P_{\boldsymbol{\eta}}(\eta) * P_{\boldsymbol{e}}(-e), \tag{6}$$

where $*$ is the convolution operator. Then, a *sufficient condition* for PEC is when one of the PMFs on the right-hand-side (RHS) of (6) is *sparse* and the other is *bounded*. As shown later, in CMOS circuits, it is quite convenient to design the main block to generate a sparse $P_{\boldsymbol{\eta}}(\eta)$ and an estimator which

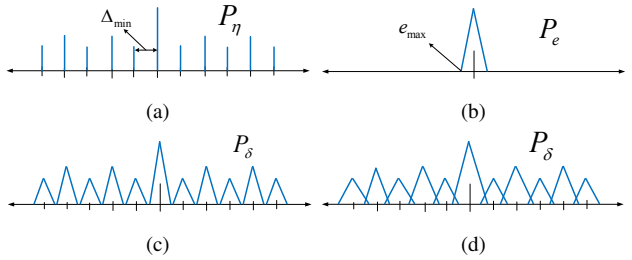


Fig. 3: Error statistics for PEC: (a) $P_{\boldsymbol{\eta}}(\eta)$, (b) $P_{\boldsymbol{e}}(e)$, (c) $P_{\boldsymbol{\delta}}(\delta)$ when the condition for PEC (8) is satisfied, and (d) $P_{\boldsymbol{\delta}}(\delta)$ when (8) is violated.

results in a bounded $P_{\boldsymbol{e}}(e)$. In this case, (6) can be expressed as:

$$P_{\boldsymbol{\delta}}(\delta) = \sum_{i=1}^{N} P_{\boldsymbol{\eta}}(\eta_i) P_{\boldsymbol{e}}(\eta_i - \delta), \tag{7}$$

where N is the number of elements in the sample set $\Omega_\eta = \{\eta_1, \eta_2, \eta_3, ..., \eta_N\}$, and the condition for PEC (see Fig. 3) is given by

$$\Delta_{\min} > 2e_{\max}, \tag{8}$$

where

$$\Delta_{\min} = \min |\eta_i - \eta_j|,\ \forall \eta_i, \eta_j \in \Omega_{\boldsymbol{\eta}}, \tag{9a}$$
$$e_{\max} = \max |e|,\ e \in \Omega_{\boldsymbol{e}}. \tag{9b}$$

Here $\Delta_{\min}$ and $e_{\max}$ provide a measure of the sparsity of $P_{\boldsymbol{\eta}}(\eta)$ and the boundedness of $P_{\boldsymbol{e}}(e)$, respectively. It is clear from Fig. 3(c) that when (8) is satisfied, the N-terms in (6) are non-overlapping. Hence, the decision rule

$$\hat{\eta} = \arg \min_{\eta_k \in \Omega_{\boldsymbol{\eta}}} |\delta - \eta_k|, \tag{10}$$

will result in $Pr\{\eta = \hat{\eta}\} = 1$ and therefore $p_\epsilon = 0$. The decision rule (10) can also be shown to be the optimal ML estimation rule for this problem (i.e, $\hat{\eta} = \arg\min_{\eta_k \in \Omega_{\boldsymbol{\eta}}} P_{\boldsymbol{\delta}|\boldsymbol{\eta}}(\delta|\eta_k)$). Note that the decision rule in (10) is a simple round off operation, which can be implemented by a truncated adder, and therefore has minimal overhead.

Note that, if (8) is not satisfied, the post-compensation error rate p_ϵ is bounded by $P(|e| > \Delta_{\min}/2)$, which is still much smaller than p_η and p_e. Thus, AEC can be used even when its post compensation error rate is non-zero similar to the other SEC techniques. In this paper, we consider AEC operation in the regime where PEC (8) is satisfied.

3.3. Application to Timing Error Cancellation

Past work [9] has shown that errors occur in the MSBs in LSB-first architectures under VOS. Thus, via an appropriate choice of the clock frequency f_{clk} and supply voltage V_{dd}, it is possible to restrict the errors in a M-bit main block to a fixed number of K MSBs. In this case, $P_{\boldsymbol{\eta}}(\eta)$ will be sparse with $\Delta_{\min} = 2^{M-K}$. For example, in an 8-bit main block with timing violations restricted to 2 MSBs, the sample set $\Omega_\eta = \{-192, -128, -64, 0, 64, 128, 192\}$ with $\Delta_{\min} = 2^{8-2} = 64$. Reduced precision replica (RPR) estimators [10] employ a reduced precision version of the main block. Thus, the estimation error e is equal to the quantization noise and thus is bounded. The rest of this paper assumes hardware errors η are due to VOS timing violations and estimation error e is due to quantization from a reduced precision estimator.

4. SIMULATION RESULTS

We present simulation results of the proposed AEC technique applied to a 45-tap low pass FIR filter in a $45\,\mathrm{nm}$ CMOS process. Similar results were obtained for a 16-tap Euclidean distance metric computation as well, but are not described here due to space limitations.

4.1. Kernel Architectures

In order to evaluate the effectiveness of AEC, we compare an AEC-based 45-tap FIR with: 1) the uncompensated FIR (identical to the main block in the AEC-based FIR), and 2) a fast (Wallace-tree multiplier and Kogge-Stone adder) main block architecture. The reason for the comparison in 2) is because a common approach to reduce energy is to voltage scale a fast architecture. Such architectures enable a greater reduction of the supply voltage V_{dd} at the same f_{clk}, and result in greater energy savings even when their total switching capacitance is larger than the slow architecture.

The main block for all three architectures (AEC, uncompensated, fast) is a 45-tap direct form FIR with a $10\,\mathrm{b}$ input, $8\,\mathrm{b}$ coefficient, and $22\,\mathrm{b}$ output, in order to meet the design requirements for a feature extractor in an EEG seizure detection system [11]. This filter employs Baugh-Wooley multipliers, a carry save adder (CSA) tree adder, and a ripple carry adder (RCA) vector merging stage. For this FIR filter, the critical voltage-frequency pair was found to be ($V_{dd-crit} = 1.2\,\mathrm{V}$, $f_{clk-crit} = 452.3\,\mathrm{MHz}$), i.e., boundary conditions for error-free operation to occur. For the AEC, a reduced precision replica (RPR) estimator [10] with a $5\,\mathrm{b}$ input, $4\,\mathrm{b}$ coefficient, and $13\,\mathrm{b}$ output, was employed. A constant is subtracted from the RPR output in order to reduce effective $e_{\max}$. It can be shown that $e_{\max} < 2^{17}$, and hence from (8), PEC is achieved when $\Delta_{\min} \geq 2^{18}$, implying that $22-18 = 4$ MSBs are permitted to be in error. This estimator incurs a complexity overhead of 34% logical area overhead. The estimator has smaller critical path and is operated in regions where it does not make hardware error.

4.2. Evaluation Methodology

Figure 4 shows the methodology employed to evaluate the system level error compensation capability of AEC and the

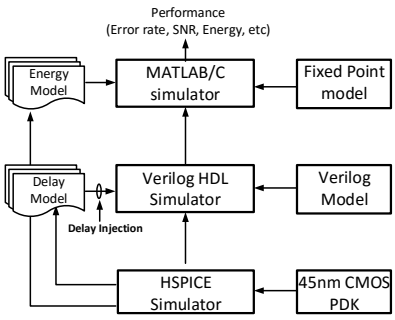


Fig. 4: Evaluation methodology.

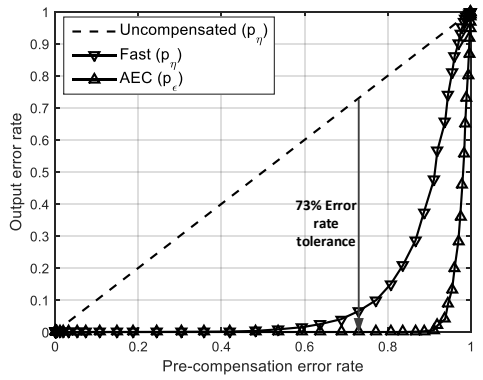
potential energy savings over the uncompensated and fast architectures. The functional behavior of the three architectures were obtained via structural Verilog simulations in which the architectures were described at the gate level. The delay T_p vs. V_{dd} characterization of each gate was done in HSPICE for $0.6\,\mathrm{V} \leq V_{dd} \leq 1.2\,\mathrm{V}$. The Verilog simulations were conducted at $f_{clk} = f_{clk-crit} = 452.3\,\mathrm{MHz}$ and for $V_{dd} \leq 1.2\,\mathrm{V}$ by assigning voltage-specific delay values to each gate. Thus, timing errors manifest due to critical path delay violations in Verilog simulations.The following energy model [12]

$$E_T = C_T V_{dd}^2 + V_{dd} I_{off} \frac{1}{f_{clk}}, \tag{11}$$

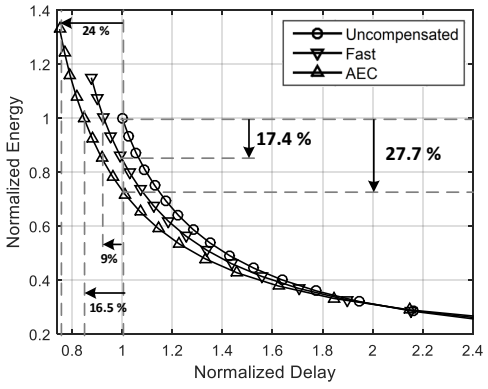
was employed to estimate energy savings, where E_T is the total energy including both dynamic and leakage, C_T is the effective load capacitance, and I_{off} is the total leakage current. This model was validated for a 20-stage RCA and its accuracy was found to be within 5% of HSPICE simulations.

4.3. Results

The AEC, the uncompensated, and the fast architecture, were simulated at $0.6\,\mathrm{V} \leq V_{dd} \leq 1.2\,\mathrm{V}$ and $f_{clk} = f_{clk-crit} = 452.3\,\mathrm{MHz}$ with 10^6 input vectors at each V_{dd}. Figure 5(a) shows that the output error rate for AEC $p_\epsilon = 0$ for the pre-compensation error rate $p_\eta \leq 0.73$ for $V_{dd} \geq 0.88\,\mathrm{V}$. This implies that AEC is able to correct errors perfectly even when the main block (also the uncompensated architecture) and the estimator are making errors 73% and 98% of the time, respectively. In contrast, the fast architecture operates error free for $V_{dd} \geq 0.98\,\mathrm{V}$. Thus, under PEC, the AEC can be viewed as a speed-up technique, as it has transformed a slow (uncompensated) architecture into one that can operate (error-free) at the same f_{clk} but at a lower voltage. This speed-up is 31.5% at the same supply voltage of $V_{dd} = 1.2\,\mathrm{V}$, and 19.7% at the same energy level (see Fig. 5(b)). Furthermore, the AEC architecture, i.e., the speed-enhanced slow architecture, is even faster than the conventional fast architecture by about 8.9% at the same energy level. At $f_{clk} = 452.3\,\mathrm{MHz}$, AEC and fast architectures result in 27.7% and 17.4% energy savings, respectively, in comparison to the uncompensated architecture. AEC exhibits energy savings of 12.4% as compared to the



(a)



(b)

Fig. 5: Simulation results comparing: (a) output error rate vs. pre-compensation error rate of the uncompensated architecture, and (b) energy vs. delay for the AEC, uncompensated, and fast FIR architectures. The output error rate of the fast architecture is obtained at the same supply voltage as the uncompensated architecture. The estimator error rate $p_e = 0.98$ when it makes no hardware errors.

fast architecture at the same frequency. Thus, as mentioned earlier, the PEC property of AEC allows comparisons with conventional energy reduction/speed-up techniques.

5. CONCLUSIONS

In this paper, we propose a novel error-resiliency technique called AEC, that established specific conditions on the error statistics of its components to achieve perfect error compensation. The proposed technique is verified on a 45-tap low pass FIR filter in a $45\,\mathrm{nm}$ CMOS process. Note that though (8) guarantees $p_\epsilon = 0$ in theory, it is difficult in practice to prove that p_ϵ is indeed zero. This problem is no different than proving that a specific architecture will indeed operate error-free at a given V_{dd} and f_{clk}. This is because of the well-known difficulty in accurately estimating the probability of highly unlikely events. The AEC technique can be applied to a variety of computational kernels, and its effectiveness in the non-PEC scenario can be studied.

6. REFERENCES

[1] A. P. Chandrakasan and R. W. Brodersen, "Minimizing power consumption in digital cmos circuits," *Proceedings of the IEEE*, vol. 83, no. 4, pp. 498–523, 1995.

[2] R. G. Dreslinski, M. Wieckowski, D. Blaauw, D. Sylvester, and T. Mudge, "Near-threshold computing: Reclaiming moore's law through energy efficient integrated circuits," *Proceedings of the IEEE*, vol. 98, no. 2, pp. 253–266, 2010.

[3] R. Hegde and N. R. Shanbhag, "A voltage overscaled low-power digital filter ic," *Solid-State Circuits, IEEE Journal of*, vol. 39, no. 2, pp. 388–391, 2004.

[4] D. Ernst, N. S. Kim, and et al., "RAZOR: a low-power pipeline based on circuit-level timing speculation," in *Microarchitecture, 2003. MICRO-36. Proceedings. 36th Annual IEEE/ACM International Symposium on*, Dec 2003, pp. 7–18.

[5] J. Tschanz, K. Bowman, S.-L. Lu, P. Aseron, M. Khellah, A. Raychowdhury, B. Geuskens, C. Tokunaga, C. Wilkerson, T. Karnik, and V. De, "A 45nm resilient and adaptive microprocessor core for dynamic variation tolerance," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2010 IEEE International*, Feb 2010, pp. 282–283.

[6] N. R. Shanbhag, R. A. Abdallah, R. Kumar, and D. L. Jones, "Stochastic computation," in *Proceedings of the 47th Design Automation Conference*. ACM, 2010, pp. 859–864.

[7] B. Shim, "Error tolerent digital signal processing," *Ph.D. dissertation , University of Illinois at Urbana-Champaign*, 2005.

[8] R. Abdallah and N. Shanbhag, "An Energy-Efficient ECG Processor in 45-nm CMOS Using Statistical Error Compensation," *Solid-State Circuits, IEEE Journal of*, vol. 48, no. 11, pp. 2882–2893, Nov 2013.

[9] Y. Liu, T. Zhang, and K. Parhi, "Computation error analysis in digital signal processing systems with overscaled supply voltage," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 18, no. 4, pp. 517–526, April 2010.

[10] B. Shim, S. R. Sridhara, and N. R. Shanbhag, "Reliable low-power digital signal processing via reduced precision redundancy," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 12, no. 5, pp. 497–510, 2004.

[11] N. Verma, K. H. Lee, and A. Shoeb, "Data-driven approaches for computation in intelligent biomedical devices: A case study of eeg monitoring for chronic seizure detection," *Journal of Low Power Electronics and Applications*, vol. 1, no. 1, pp. 150–174, 2011.

[12] S. Zhang and N. R. Shanbhag, "Reduced overhead error compensation for energy efficient machine learning kernels," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*. IEEE Press, 2015, pp. 15–21.